木構造属性を許容する決定木学習*

1C－6

秋葉 泰弘†　フセイン アルモアリム‡　金田 重郎†
NTTコニュニケーション科学研究所†　サウジアラビア国立石油鉱物大学‡

1 始めに

本論文では，事例を表現する属性が木構造を有する場合の決定木学習を取り上げる．従来手法としてQuinlanによるエンコーディング・アプローチがあるが，現実の問題で取り扱う大きな木構造の場合は，計算量や未知事例に対する正解率の点で問題があった．提案手法は，エンコーディングによるアプローチとは違い，木構造属性を直接取り扱え，事例の前処理を必要としない．提案手法と前述のQuinlanのアプローチの性能を比較するために，現実データと人工データで実験をしたところ，提案手法の方が，２倍～４倍計算時間が早く，未知事例に対して高い正解率を示した．

2 タスク

木構造属性は，離散値をとる属性の一種である．その取り得る値は単なる値のリストではなく，is-a関係からなる階層構造を成す．以下，この階層構造を**属性木**と呼ぶ．例えば，事例がそれらの形と色で記述される対象領域を考えよう．図1の木は属性“色”の属性木を示す．この木の各エッジは，あるノードとその親との間のis-a関係を表現している．これらの木のノード上の値を，**カテゴリー**と呼ぶ．

木構造属性を許容する決定木学習とは，事例を属性木に沿って汎化を行ない，事例の表現より高いレベルのカテゴリーでルールを表現可能な決定木学習である．各節における質問は，“ある木構造属性xに対して，その属性値は，カテゴリーvまたはその属性木上の下位カテゴリーか?”なる質問である．以下，この質問を“$x \equiv v$か?”と表記する．

木構造属性を許容する決定木学習は，より高いレベルのカテゴリーでルールを表現されるため，より端的な方法で概念を表現可能となり，それ故に，汎化性能がよくなる．

一方，学習アルゴリズムが，属性木にアクセスしなければ，その出力は，それらの値(カテゴリー)だけを利用して，表現されねばならない．その結果，事例のアルゴリズムがこれらの訓練事例を汎化する能力は，制限される．また，大量の訓練事例が準備されない限り，貧相な予測性能しか示せない．

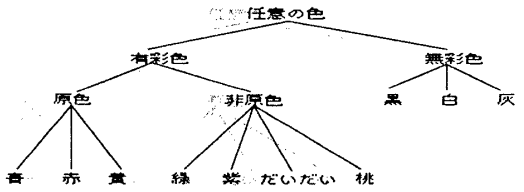

図1: 属性“色”の属性木

3 提案手法

決定木学習における基本となるタスクは，与えられた訓練事例の集合Sに対して，一番いいスコアを持つ質問を見つけることである．木構造属性の場合には，与えられた木構造属性xと与えられた訓練事例Sに対して，xの全てのカテゴリーの中で，“$x \equiv v$か?”なる質問に対するスコアが一番高いカテゴリーvを探すことになる．与えられた属性xに対して，その様なvを見つけることを，以後**属性xを処理する**と言う．

著者らの方法を紹介する前に，用語を定義する．$e \in S$を訓練事例とし，xをその属性木がT_xの木構造属性とする．そして，eの属性xの値がvであるとし，PがT_xのルートからノードvへ至るパスであるとする．この時，**事例eをT_xに沿って流す**という処理は，パスP上のノードをルートから順に下ってvに至るまで辿ることをいう．

木構造属性xに対して，その属性木T_xの各ノードvに各クラス用のエントリーを持つ配列CF_vを準備する．この配列は，質問“$x \equiv v$か?”を満足する事例からなる部分集合におけるクラス頻度が格納される．vがT_xのルートであれば，CF_vは，丁度Sのクラス頻度になる．本稿では，これを**基本クラス頻度**と呼ぶ．

以下，与えられた属性を処理する手続きを概説する．まず，各事例$e(\in S)$をT_xに沿って流す．事例eを流した際にノードvを経由したとすれば，エントリー$CF_v[c]$は一つカウントアップされる．ここで，cは，事例eのクラスである．全ての事例を流すと，各配列CF_vは$x \equiv v$を満足する事例の部分集合中のクラス頻度を表現して

*Decision Tree Learner Handling Tree-Structured Attributes,
Yasuhiro Akiba †, Hussein Almuallim ‡, Shigeo Kaneda †
†)NTT Communication Science Laboratories, 1-2356, Take, Yokosuka-shi, Kanagawa-ken, 238-03, JAPAN
‡) the Dept. of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261,Saudi Arabia

いる．残りの事例 ($x \not\equiv v$ を満足する事例) からなる部分集合のクラス頻度は，基本クラス頻度と CF_v の差である．これらの頻度が計算されると，属性木中の任意のノード v に対して，質問 "$x \equiv v$ か?" のスコアの計算が可能になる．以下の点を考慮すれば，よりよいインプリメンテーションが可能になる．：

- 木構造属性が処理される度に，各ノードに対するクラス頻度配列を明示的に初期化するのは避けるべきである．大域変数 *CurrentTime* を管理すること及び各ノード v 毎に整数値変数 *LastAccessTime* を準備することにより回避できる．変数 *CurrentTime* は，木構造属性が処理される度に，処理の始めに一度だけカウントアップされる．ある事例 e を流し，あるノード v に到達する度に，その *LastAccessTime* と *CurrentTime* を比較する．もし等しければ，上述の様に CF_v の適切なエントリをカウントアップすることにより，通常通り処理される．しかし，*LastAccessTime* が，*CurrentTime* より小さければ，まず，それを *CurrentTime* と等しく設定し，次いで CF_v の各エントリーを 0 に初期化し，最後に CF_v の適切なエントリを 1 増加させる．

- 各可能な質問に対するスコアを計算する際 (全ての事例が流され，クラス頻度が既に計算されている)，カテゴリーのいくつかは安全に無視できる．というのは，それらのスコアが最善でないことが事前に決定され得るからである．この点は，次の 2 つの場合に成り立つ．：

  1. v を，全ての事例を流した後に *LastAccessTime* が *CurrentTime* より小さいノードは，質問 "$x \equiv v$ か?" のスコアが，0 である．従って，このノード及びその子孫は，無視すべきである．

  2. v を，CF_v のエントリーが，一つのエントリを除いて 0 であるノードであるとする．この場合，v の子孫は，v よりよいスコアは持てない．従って，安全に無視できる．

  実行すべき処理は，上述のタイプ (1) のノードとその子孫，及びタイプ (2) の全てのノードの子孫を無視した，縦型探索である．

4 実験的比較

この章では，Quinlan エンコーディングアプローチと著者らの直接的アプローチとを実験的に比較する．この実験では，日本語動詞を英語動詞に対応させるルールを学習する際の，2 つの手法の汎化性能を比較した．

この対象領域では，訓練事例は，⟨ J-sentence, E-verb ⟩ なる形式である．日本語文は前処理され，主格や目的格の様な文の構成要素のベクトルで表現される．各成分は，NTT で開発中の日英機械翻訳システム ALT-J/E 中の意味シソーラス上のカテゴリーである．この意味シソーラスは，約 2,700 個のカテゴリーを持ち，深さが 12 段である．従って，主格，目的格等は，属性木が ALT-J/E の意味シソーラスである，木構造属性である．

表 1: 直接法 (Direct) と Quinlan-encoding 法 (QE) で獲得された英語動詞選択ルールのエラー率 (%)

日本語動詞	英語動詞数	事例数	枝刈なし	
			Direct	QE
行なう	3	31	5.8	47.5
解く	3	28	16.7	6.7
使う	3	45	0.0	4.0
焼く	5	27	18.3	70.0
作る	36	167	45.4	53.3
飲む	10	159	19.4	47.2
引く	20	94	55.5	53.2
平均			24.9	38.3

実験は，異なる 12 種類の日本語動詞で行なった (表 1を参照)．各々は，3 個から 36 個の対訳英語動詞を持つ．ここで考察している 2 つの手法の汎化性能を評価するために，各日本語動詞毎に，事例に対して 10-fold cross valication を実行した．表 1が示すように，この対象領域では，直接的アプローチの方が，Quinlan エンコーディングアプローチよりよく振舞う．また，直接的方法は，Quinlan エンコーディング法より平均で 2 から 4 倍早く走った．

5 おわりに

本稿では，対象領域の属性が木構造属性である時に，決定木を学習する問題を議論した．この問題を解決するために，訓練事例に前処理を必要とせず，直接木構造属性を取り扱うアプローチを導入した．直接的アプローチを Quinlan エンコーディングと実験的に比較した．直接的アプローチの方が，約 2 倍から 4 倍速いことが分かった．また，直接的アプローチは，Quinlan エンコーディングより良い汎化性能を示すことを実験で確認した．しかし，これら 2 つの方法は，異なるバイアスの実行であり，この様な言及は一般的には成り立たないであろう．

ここで提案した方法は，属性が is-a 階層構造が木構造ではなしに，DAG (directed acyclic graph) をなす場合に容易に拡張出来る．

参考文献

[1] 池原，宮崎，横尾，" 日本語機械翻訳のための意味解析辞書 "，電子情報通信学会，研究会報告，NLC 91-19,(1991).

[2] Quinlan, J. R. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann 1993.